GHAJAR
EXHIBIT 21

UNITED STATES DISTRICT COURT
NORTHERN DISTRICT OF CALIFORNIA
SAN FRANCISCO DIVISION

RICHARD KADREY, AN	: CASE NO.
INDIVIDUAL; SARAH SILVERMAN,	: 3:23-cv-03417-VC
AN INDIVIDUAL; CHRISTOPHER	:
GOLDEN, AN INDIVIDUAL,	:
PLAINTIFFS	:
VS.	:
META PLATFORMS, INC., A	:
DELAWARE CORPORATION,	:
DEFENDANT	:
________________________________	:

HIGHLY CONFIDENTIAL - ATTORNEYS' EYES ONLY

VIDEOTAPED DEPOSITION OF JONATHAN KREIN, PH.D.
SAN FRANCISCO, CALIFORNIA
THURSDAY, MARCH 6, 2025

REPORTED BY:
DEBBIE LEONARD, CSR, RDR, CRR
CSR NO. 14350
JOB No. 7189213

PAGES 1 to 113

Page 1

MR. YOUNG: Objection. Form.

THE WITNESS: So, for example, in paragraph 81, I discuss deduplication, lay it out there.

So the substance of the opinions that I've formed with respect to deduplication is around its impact on the generation of copies.

Exactly what this particular instance of deduplication was used for by Meta was not something I was asked to consider.

BY MR. WEINSTEIN:

Q Understood.

If you could turn to paragraph 97. You identified as part of your analysis certain Python scripts that would remove certain information from the content in The Pile dataset, correct?

A Could you repeat the question, please?

Q Yes, sir.

You identified in paragraph 97 a Python script that removes information from various aspects of the The Pile dataset, correct?

A Yes.

Q Okay. Now, the Python script that you identified, pile_clean_v0.py, that is not limited to Books3; it also applies to ArXiv, A-R-X-I-V; GitHub; and other portions of Books3, correct?

Page 77

A Correct.

Q Okay. And then let's go to paragraph 99. And is this a list of key words that are removed by one of the Python scripts that you analyzed with respect to The Pile?

A So this list of terms refers to paragraph 98 of my report, which discusses a script called pile_clean_v1.py, and it is referring to terms that were used to identify lines or segments of -- of The Pile data to be removed.

Q And you have a list on paragraph 99 of key words that are used for identifying [indiscernible], correct?

(Court reporter clarification.)

BY MR. WEINSTEIN:

Q Paragraph 99, you have a list of key words for use in identifying material to be removed, correct?

MR. YOUNG: Objection. Form.

THE WITNESS: So as I just mentioned, the list in paragraph 99 refers to the script in paragraph 98, and these are key words that are for use in identifying material to be removed, yes.

BY MR. WEINSTEIN:

Q Okay. And they include "Published by," "Produced by," "Copyright," "Cover copyright," "Cover designed by," "Edited by," "Prologue," "PREFACE,"

Page 78

"Epilogue," "Acknowledgments," "About the Author," "Slashwords [sic] Edition," "All rights reserved," "Title Page," "Table of Contents," "Introduction," "Cover," "Title Page," and again the word "Copyright."

Do you see that?

A I do see that, with the exception that I believe you may have said "Slashwords Edition," and it says "Smashwords Edition."

Q Thank you.

MR. WEINSTEIN: What's next in order?

THE REPORTER: Number 5.

MR. WEINSTEIN: 5. Thank you.

I'd like to mark as Exhibit 5 a copy of a document entitled pile_clean_v1.py. That's META-KADREY-SC-000625 through 629.

(Krein Exhibit 5 marked for identification.)

BY MR. WEINSTEIN:

Q Dr. Krein, the court reporter just handed you a document that's called pile_clean_v1.py. This is one of the source code files that you looked at in connection with your opinions related to The Pile, correct?

A Correct.

Q And this is a Python file, correct?

A Yes.

Q And just humor me for the record. What does

Python refer to?

A Python is a programming language.

Q If you could turn to the page that begins 000628 and look around line 304.

A I'm on page 628. Well --

Q And you see line 304?

A I do, yes.

Q And that's the main program, correct?

A That's the main entry point.

Q Correct. So when the script is executed, that's the first part of the code that will get executed, correct?

A No.

Q Okay. Will it get executed at all?

A Yeah. To clarify, on the next page, you'll see at line 371, that's where main is called from. And then at 304, it executes.

Q I see. So main will then call main, and then the trip will function, right?

A Yeah.

Q Okay. And do you see where it says "cleaners equals," and there's a list?

A I do, yes.

Q What does that list refer to?

A So my understanding is that these would be parts

Page 80

of -- parts of The Pile dataset.

Q Are these portions of The Pile dataset for which cleaning operations are performed?

MR. YOUNG: Objection. Form.

THE WITNESS: So these are portions, as Meta has named them here, of The Pile dataset for which there are, yes, cleaning operations that get performed on any part of the data that's designated under one of these designations here.

BY MR. WEINSTEIN:

Q And there are 22 of them, correct, approximately?

A There may be. I'd have to go through and count.

Q Okay. And you see that on line 313, it says "Books3"?

Do you see that?

A Line 313 refers to "Books3."

Q Okay. Is that the cleaner that relates to Books3?

MR. YOUNG: Objection. Form.

BY MR. WEINSTEIN:

Q Or it's identifying the cleaner that relates to Books3?

A So that is a text designation that points to the name of a function which is called "book3" at line 63.

When the parameter "Books3" is provided, then it results in the function at 63 being executed.

Q Understood.

We can go back to your report now, sir, Exhibit 1. And if you could turn to paragraph 100, this is another file you identified, a .scala file, that removes data from the Books3 dataset, correct?

A So paragraph 100 of my report: "BooksHeaderCleaner.scala removes copyright information from the Books3 dataset."

Q And just so we know, what is a Scala file?

A It is a type of programming file.

Q And then you have a long list of key words that goes from page 46 to 47 in your report. Are these key words that are removed from the Books3 dataset when this program is executed?

A So the key words are used to identify segments or lines -- I believe in this case, it's -- it may be lines; I'd have to look back -- but segments of text to be removed from the data.

Q And at the top of page 47, you mention that there are a number of other terms, like the word "facebook" being removed?

A The word "facebook" is one of the terms that is additionally used in filtering lines.

Q And "www.," correct?

A That also is one of the terms that I've listed here in paragraph 100.

Q And "notebook" is another one, correct?

A "Notebook" as a single word, and "note-book" as well.

Q Okay. Do you know why this script would be designed to remove the word "facebook" from the original text?

MR. YOUNG: Objection to form. Scope.

THE WITNESS: I did not examine that particular question and didn't form any opinion about it.

BY MR. WEINSTEIN:

Q Okay. And then on paragraph 101, you discuss another source code file called BooksFooterCleaner.scala, and you have another list of key words, correct?

A Yes, paragraph 101 does refer to another programming script that also lists a set of key words that are used to identify text.

Q And these include things like "already a subscriber," "thank you for downloading," "thank you for purchasing," among others, correct?

A To the extent that you read those out of this list, then they would, of course, be in the list.

Q Now, when you analyzed -- withdrawn.

Page 83

CERTIFICATE

I, Debbie Leonard, Certified Shorthand Reporter No. 14350 for the State of California, do hereby certify:

That the foregoing deposition was taken before me at the time and place therein set forth, at which time the witness was put under oath by me; that the testimony of the witness and all objections made at the time of the examination were recorded stenographically by me, were thereafter transcribed by me by means of computer; and that the foregoing is a true record of same.

I further certify that I am neither counsel for nor related to any party to said action, nor in any way interested in the outcome thereof.

IN WITNESS WHEREOF, I have subscribed my name this 13th day of March, 2025.

Debbie Leonard

Debbie Leonard, CSR, RDR, CRR
CSR NO. 14350